Lecture 1: Describing social networks

Short Course on Social Interactions and Networks,

CEMFI, May 26-28th, 2014

Bryan S. Graham, UC - Berkeley

26 May 2014

Figure 1 provides a visual representation of a set of risk-sharing links, measured in the year 2000, between 119 households residing in Nyakatoke, a small village in Tanzania. These data are described and analyzed by de Weerdt (2004). Individuals were asked for lists of people that they could "personally rely on for help". A list of undirected links between all households was constructed using responses to this question.

Each point in the figure represents a household, with the size of the point proportional to the number of risk sharing links to which the household is party. Yellow, orange, green and blue households correspond to categories of increasing land and livestock wealth (see the notes to Figure 1).

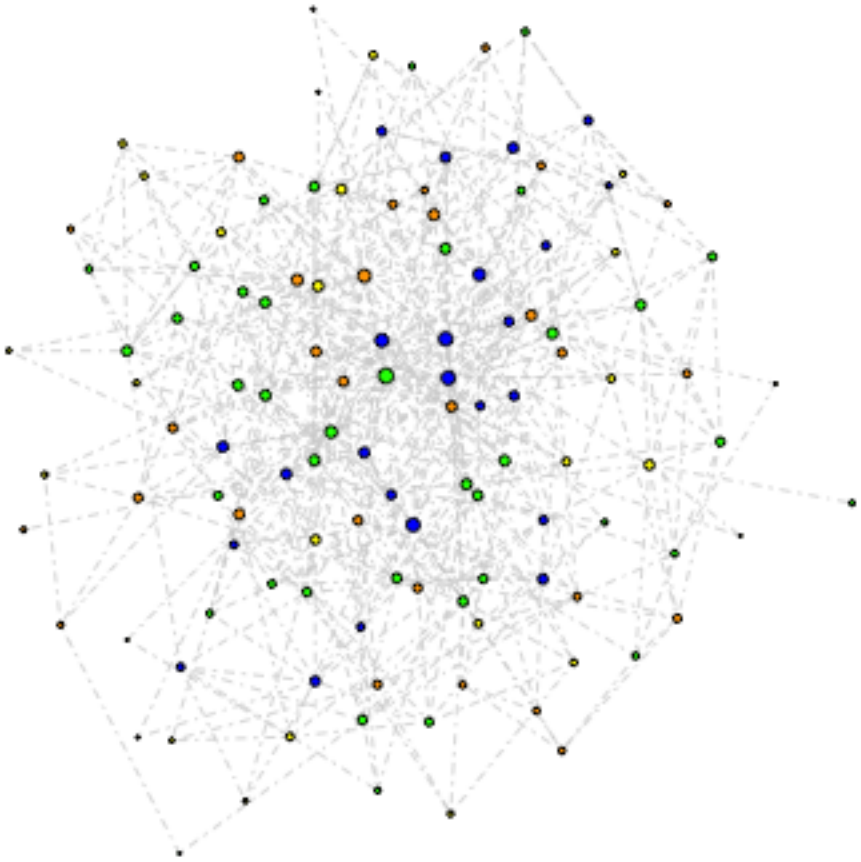
This note describes methods for summarizing network data of the type depicted in Figure 1. Basic references for the material survey below include Newman (2003), Jackson (2008) and Kolaczyk (2009). A few minor results presented below, mostly of pedagogical significance, are new.

An **undirected graph** $G(\mathcal{N},\mathcal{E})$ consists of a set of **nodes** $\mathcal{N}=\{1,\ldots,N\}$ and a list of unordered pairs of nodes called **edges** $\mathcal{E}=\{\{i,j\},\{k,l\},\ldots\}$ for $i,j,k,l\in\mathcal{N}$. A graph is conveniently represented by its **adjacency matrix** $\mathbf{D}=[D_{ij}]$ where

$$D_{ij}=\begin{cases}1 & \text{if } \{i,j\}\in\mathcal{E}\\ 0 & \text{otherwise}\end{cases}. \tag{1}$$

A node, depending on the context, may be called a vertex, agent or player. Likewise edges may be called links, friendships, connections or ties. Since self-ties are ruled-out, and the nodes in edges are unordered, the adjacency matrix is a symmetric binary matrix with a

Figure 1: Nyakatoke risk-sharing network



Source: de Weerdt (2004) and author's calculations.
Notes: Node size proportional to household degree. Yellow nodes represent households with land and livestock wealth below 150,000 Tanzanian shillings, orange those between 150,000 and 300,000, green those between 300,000 and 600,000 and blue those with 600,000 and above. Following Comola and Fafchamps (forthcoming) land was valued at 300,000 shillings per acre. Network plotted using igraph package in R (see `http://igraph.org/r/`).

diagonal of so-called structural zeros (i.e., $D_{ij} = D_{ji}$ and $D_{ii} = 0$).

A **social network** consists of a set of agents (nodes) and ties (edges) between them. A social network can be conveniently represented by it node and edge list or by its adjacency matrix. I will utilize the adjacency matrix representation in most of what follows.

In summarizing the structure of a social network it is convenient to define network statistics at the level of individual agents, at the level of pairs of agents or **dyads**, and at the level of triples of agents or **triads**.

Network statistics involving single agents

The total number of links belonging to agent i, or her **degree** is $D_{i+} = \sum_j D_{ij}$. The degree frequency distribution of a network, or **degree distribution** for short, consists of the frequency of each possible agent-level degree count $\{0, 1, \ldots, N\}$ in the network. A important component of the literature on networks takes the degree distribution as its primitive object of interest (e.g., Barabási and Albert (1999) and Albert and Barabási (2002)). This focus is motivated by the fact that many other topological features of a network are fundamentally constrained by its degree distribution. Jackson and Rogers (2007) develop connections between degree distributions and the spread of infectious diseases and new technologies/behaviors through a network. I will have more to stay about the connection between a network's degree sequence and its other topological features below.

The density of a network equals the frequency with which any randomly drawn dyad is linked:

$$P_N = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} D_{ij}. \tag{2}$$

Note that $(N-1)\, P_N$ coincides with average degree. The density of the Nyakatoke network is 0.0698.

Bonacich (1972) recursively defined an agent's **centrality,** power, or importance within a network, C_i, to be proportional to the sum of her links to other agents, weighted by their own centralities. Letting $\mathbf{c}$ be the N vector of centrality measures this gives

$$\lambda \mathbf{c} = \mathbf{D}\mathbf{c}.$$

Since $(\mathbf{D} - \lambda I_N)\, \mathbf{c} = 0$, Bonacich's measure corresponds to a normalized eigenvector of $\mathbf{D}$. Typically the largest eigenvalue $\lambda_{\max}$ is used for normalization. Jackson (2008) surveys additional measures of node centrality.

Consider the matrix product

$$\mathbf{D}^2 = \begin{pmatrix} D_{1+} & \sum_i D_{1i}D_{2i} & \cdots & \sum_i D_{1i}D_{Ni} \\ \sum_i D_{1i}D_{2i} & D_{2+} & \cdots & \sum_i D_{2i}D_{Ni} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i D_{1i}D_{Ni} & \sum_i D_{2i}D_{Ni} & \cdots & D_{N+} \end{pmatrix}.$$

The i^{th} diagonal element of $\mathbf{D}^2$ equals the number of agent i's links or her degree**.** The $\{i,j\}^{th}$ element of $\mathbf{D}^2$ gives the number of links agent i has in common with agent j (i.e., the number of "friends in common"). In the language of graph theory the $\{i,j\}^{th}$ element of $\mathbf{D}^2$ gives the number of **paths** of length two from agent i to agent j. For example, if i and j share the common friend k, then a length two path from i to j is given by $i \to k \to j$. The diagonal elements of $\mathbf{D}^2$ correspond to the number of length two paths from an agent back to herself. For example if i is connected to k, then one such a path is $i \to k \to i$. The number of such paths coincides with an agent's degree.

Calculating $\mathbf{D}^3$ yields

$$\mathbf{D}^3 = \begin{pmatrix} \sum_{i,j} D_{1i}D_{ij}D_{j1} & \sum_{i,j} D_{1i}D_{ij}D_{j2} & \cdots & \sum_{i,j} D_{1i}D_{ij}D_{jN} \\ \sum_{i,j} D_{1i}D_{ij}D_{j2} & \sum_{i,j} D_{2i}D_{ij}D_{j2} & \cdots & \sum_{i,j} D_{2i}D_{ij}D_{jN} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i,j} D_{1i}D_{ij}D_{jN} & \sum_{i,j} D_{2i}D_{ij}D_{jN} & & \sum_{i,j} D_{Ni}D_{ij}D_{jN} \end{pmatrix},$$

whose $\{i,j\}^{th}$ element givens the number of paths of length 3 from i to j.

The diagonal elements of $\mathbf{D}^3$ are counts of the number of transitive triads or **triangles** in the network. If both i and j are connected to k as well as to each other, then the $\{i,j,k\}$ triad is closed (i.e., "the friend of my friend is also my friend"). Note that if $\{i,j,k\}$ is a closed triad it is counted twice each in the i^{th}, j^{th} and k^{th} diagonal elements of $\mathbf{D}^3$. Therefore $\mathrm{Tr}\left(\mathbf{D}^3\right)/6$ equals the number of unique triangles in the network.

Proceeding inductively it is easy to show that the $\{i,j\}^{th}$ element of $\mathbf{D}^K$ gives the number of paths of length K from agent i to agent j.

Theorem 1. *The $\{i,j\}^{th}$ element of $\mathbf{D}^K$ gives the number of paths of length K from agent i to agent j.*

Proof. Let $D_{ij}^{(K)}$ denote the $\{i,j\}^{th}$ element of $\mathbf{D}^K$. Begin by observing that $\mathbf{D}^0 = I_N$, correctly implying that the only zero length walks in the network are those from each agent to herself. Under the maintained hypothesis, $D_{ij}^{(K)}$ equals the number of K-length paths

from i to j. The number of $K+1$ length paths from i to j then equals

$$\sum_{k=1}^{N} D_{ik}^{(K)} D_{kj},$$

which equals the $\{i,j\}^{th}$ element of $\mathbf{D}^{K+1}$. The claim follows by induction. □

Network statistics involving pairs of agents or dyads

The **distance** between agents i and j corresponds to the minimum length path connecting them. If there is no path connecting i to j, then the distance between them is infinite. We can use powers of the adjacency matrix to calculate these distances. Specifically,

$$M_{ij} = \min_{k} \left\{ k : D_{ij}^{(k)} > 0 \right\}$$

equals the distance from i to j (if it is finite). For modest sized networks M_{ij} can be calculated by taking successive powers of the adjacency matrix.

If the network consists of a single, giant, connected component, such that the minimum length path between any two agents is finite, we can compute average path length as

$$\overline{M} = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} M_{ij}. \tag{3}$$

If the network consists of multiple connected components, standard practice is to compute average path length within the largest one. Alternatively, following Newman (2003), we can calculate average distance or path length in the network as

$$\overline{M}^{\text{alt}} = \left[\binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} M_{ij}^{-1} \right]^{-1}. \tag{4}$$

By taking the reciprocal of an average of reciprocal distances, we neatly handle the 'problem' of infinite paths.

The **diameter** of a network is the largest distance between two agents in it. It will be finite if the network consists of a single connected component (in which case all agents are "reachable" starting from any given agent) and infinite in networks consisting of multiple components (in which case there are no paths connecting some pairs of agents).

Table 1 gives the frequence of minimum path lengths in the Nykatoke network. There are 490 direct ties in the network (paths of length one). Just under 7 percent of all *pairs* of

Table 1: Frequency of degrees of separation in the Nyakatoke network

	1	2	3	4	5
Count	490	2666	3298	557	10
Frequency	0.0698	0.3797	0.4697	0.0793	0.0014

Source: de Weerdt (2004) and author's calculations.

Table 2: Nyakatoke risk-sharing network triad census

	empty	**one-edge**	**two-star**	**triangle**
Count	221,189	48,245	4,070	315
Proportion	0.8078	0.1762	0.0149	0.0012
Random Graph Proportion	0.8049	0.1812	0.0136	0.0003

Source: de Weerdt (2004) and author's calculations.
Notes: The Nyakatoke network includes $N = 119$ households, corresponding to $\binom{N}{2} = 7,021$ unique dyads and $\binom{N}{3} = 273,819$ unique triads.

households are directly connected in Nykatoke. Another 2,666 dyads are only two degrees apart. That is, although they are not connected directly, they share a tie in common. About 80 percent of dyads are separated by three or fewer degrees. The diameter of the Nyakatoke network is 5. The juxtaposition of low density (i.e., only a small fraction of all possible ties exists), with few degrees of separation (i.e., low average degree and/or diameter) is a feature of many real world social networks.

The analysis of distances and diameter has a long history in social network analysis and falls under the rubric of the "small-world problem". Stanley Milgram (1967) popularized this phrase and, through a series of postal experiments in the 1960s, showed that two random individuals in the United States could be often be connected through a short chain of acquaintances ("six degrees of separation").

Network statistics involving triples of agents of triads

Triads, a set of three unique agents, come in four types: no connections, one connection, two connections, or three connections between them. These triad types are called **empties**, **one-edges**, **two-stars** and **triangles** respectively. There are $\binom{N}{3} = \frac{N!}{3!(N-3)!} = \frac{N(N-1)(N-2)}{6}$ unique triads in a network of size N. A complete enumeration of them into their four possible types constitutes a triad census.[1]

Each agent can belong to as many as $(N-1)(N-2)$ triangles. The counts of these triangles

[1]Note that dyads are either linked or not, hence the dyad census coincides with computing the total number of links in the network.

are contained in the N diagonal elements of $\mathbf{D}^3$. However each such triangle appears 6 times in these counts: as $\{i,j,k\}$, $\{i,k,j\}$, $\{j,i,k\}$, $\{j,k,i\}$, $\{k,i,j\}$ and $\{k,j,i\}$. Thus

$$\#\text{ of triangles} = T_T = \frac{\text{Tr}\left(\mathbf{D}^3\right)}{6} \tag{5}$$

equals the number of unique triangles in the network.

Each pair of agents $\{i,j\}$ can share of up to $N-2$ links in common. If $\{i,j\}$ are not linked themselves, then they may belong to two stars with their links in common as star centers. Since each dyad can create up to $N-2$ closed triangles by forming a link between themselves, there may be up to $\binom{N}{2}(N-2) = \frac{N(N-1)(N-2)}{2}$ (actual) triangles or two stars (i.e., potential triangles) in the network. These counts are contained in the lower (or upper) off-diagonal elements of $\mathbf{D}^2$. Each triad appears three times in these counts: as $\{i,j,k\}$, $\{i,k,j\}$ and $\{j,k,i\}$. If the triad is a two star, then only one of $D_{ji}D_{ki}$, $D_{ij}D_{kj}$, or $D_{ik}D_{jk}$ quantities will equal one (i.e., contribute). If it is a triangle, then all three will equal one. Therefore $\text{vech}\left(\mathbf{D}^2\right)'\iota$ gives the network count of *three times* the number triangles *plus* the number of two-stars, with the count of the latter alone equal to

$$\#\text{ of two stars} = T_{TS} = \text{vech}\left(\mathbf{D}^2\right)'\iota - \frac{\text{Tr}\left(\mathbf{D}^3\right)}{2}. \tag{6}$$

We can use a similar logic to calculate the number of one-edge triads. Each agent belongs $N-2$ triads. If all triads are empty or have only one edge, then there will be $(N-2)\,\text{vech}\left(\mathbf{D}\right)\iota$ one edge triads. However if some triads are two-stars or triangles this count will be incorrect. It turns out that subtracting twice the number of two stars and three times the number of triangles gives the correct answer.

$$\#\text{ of one edges} = T_{OE} = (N-2)\,\text{vech}\left(\mathbf{D}\right)'\iota - 2\text{vech}\left(\mathbf{D}^2\right)'\iota + \frac{\text{Tr}\left(\mathbf{D}^3\right)}{2} \tag{7}$$

The number of empty triads, T_E, equals $\binom{N}{3}$ minus the sum of (5), (6) and (7). Note that (5), (6) and (7) collectively imply that

$$\begin{aligned} T_{OE} + 2T_{TS} + 3T_T &= (N-2)\,\text{vech}\left(\mathbf{D}\right)'\iota, \\ &= \frac{1}{4}N\left(N-1\right)\left(N-2\right)P_N \end{aligned}$$

suggesting that network density can be computed from the triad census according to

$$P_N = \left(\frac{4T_{OE} + 8T_{TS} + 12T_T}{N\left(N-1\right)\left(N-2\right)}\right). \tag{8}$$

The triad census for the Nyakatoke network is given in Table 2. As a point of comparison the proportion of each type of triad that we would expect to see in a random graph, where the probability of a link between any two agents coincides with the observed density of the Nyakatoke network (0.0698), is given in the last row of the table.

A measure of network **transitivity** is given by three times the number of transitive triads in the network relative to three times the number of transitive triads *plus* those triads which could become transitive with the addition of a single link (i.e., two stars). The **Transitivity Index**, sometimes called the clustering coefficient, is

$$\begin{aligned} \text{Transitivity Index} &= \frac{3T_T}{T_{TS} + 3T_T} \\ &= \frac{1}{2}\frac{\text{Tr}\left(\mathbf{D}^3\right)}{\text{vech}\left(\mathbf{D}^2\right)'\iota} \\ &= R_N. \end{aligned}$$

In random graphs R_N should be close to network density. For the Nyakatoke network the transitivity index is 0.1884, which substantially exceeds the density of the network (0.0698). We will explore how to assess the statistical significance of this difference in a later lecture.

Transitivity has been hypothesized to facilitate risk sharing and other activities where monitoring may be helpful. If the (i, j, k) triad is transitivity, then agent k may be able to monitor actions involving i and j. See Jackson (2014) for additional discussion.

Degree distributions

Barabási and Albert (1999) assert that the degree distribution of a network, at least over some range, is well-described by a discrete Pareto or 'power law' distribution:

$$F\left(d_+\right) = 1 - \frac{A}{1-\alpha} d_+^{1-\alpha}$$

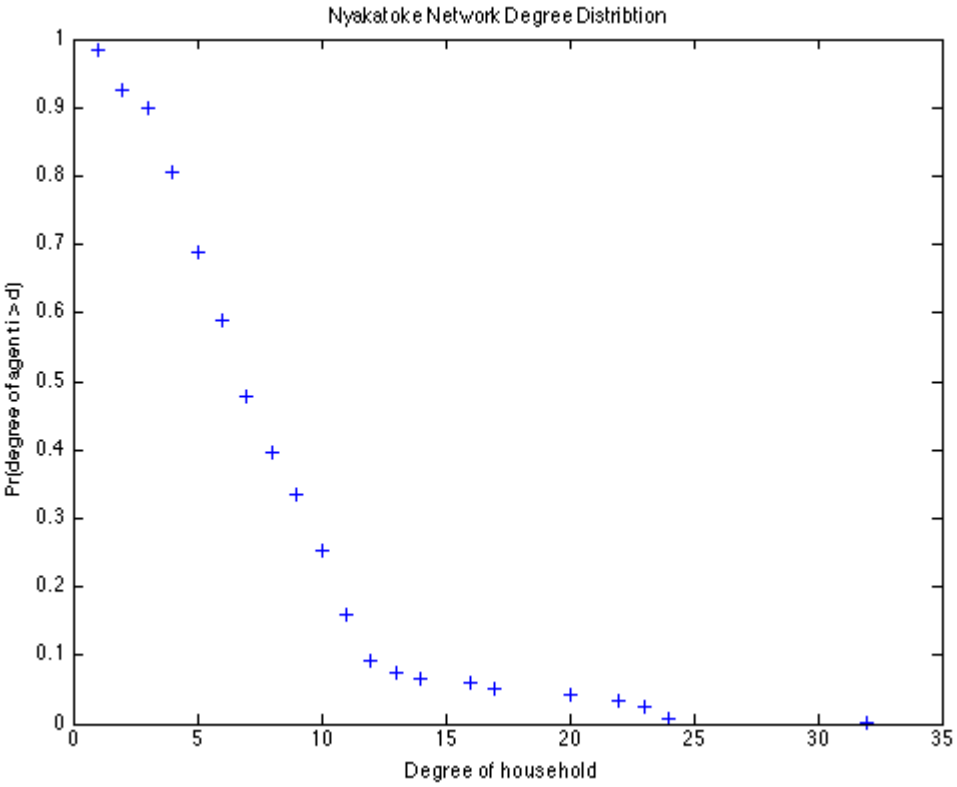
for $d_+ = \underline{d}_+, \ldots, N$ and $F\left(d_+\right) = \Pr\left(D_{i+} \leq d_+\right)$. Here $\underline{d}_+ > 0$ is some threshold degree level below which the power law distribution may not apply.

Taking logs yields the linear relationship

$$\ln\left(1 - F\left(D_{i+}\right)\right) = \ln\left(\frac{A}{1-\alpha}\right) + (1-\alpha)\ln D_{i+}.$$

The coefficient, $1-\alpha$, may be estimated by OLS (although this may not be the best approach in practice). For a review of estimation methods for power law coefficients see Clauset, Shalizi

Figure 2: Nyakatoke risk-sharing network degree distribution



Source: de Weerdt (2004) and author's calculations.

and Newman (2009).

Figure 2 plots the Nyakatoke network's degree distribution. A small number of households in the Nyakatoke network have many links (over 20), while the vast majority have only a small number of links (less than 10).

The variance of the degree distribution equals

$$S_N^2 = \frac{2}{N}\left(T_{TS} + 3T_T\right) - (N-1)\,P_N\left[1 - (N-1)\,P_N\right]. \tag{9}$$

To see this observe that

$$\begin{aligned}
S_N^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(D_{i+}-\frac{1}{N}\sum_{j=1}^{N}D_{j+}\right)^2 \\
&= \frac{1}{N}\left[\left(\mathbf{D}-\frac{\iota'\mathbf{D}\iota}{N}\right)\iota\right]'\left[\left(\mathbf{D}-\frac{\iota'\mathbf{D}\iota}{N}\right)\iota\right] \\
&= \frac{1}{N}\left[\iota'\mathbf{D}^2\iota-\iota'\mathbf{D}\frac{\iota'\mathbf{D}\iota}{N}\iota-\iota'\frac{\iota'\mathbf{D}\iota}{N}\mathbf{D}\iota+\left(\frac{\iota'\mathbf{D}\iota}{N}\right)^2\right] \\
&= \frac{\iota'\mathbf{D}^2\iota}{N}-\left(\frac{\iota'\mathbf{D}\iota}{N}\right)^2 \\
&= \frac{2\text{vech}\left(\mathbf{D^2}\right)'\iota}{N}+\frac{\iota'\mathbf{D}\iota}{N}-\left(\frac{\iota'\mathbf{D}\iota}{N}\right)^2 \\
&= \frac{2}{N}\left(T_{TS}+3T_T\right)-(N-1)P_N\left[1-(N-1)P_N\right].
\end{aligned}$$

Consider the effect of inducing a mean preserving spread in a network's degree distribution. That is, we seek manipulations which keep network density fixed, while increasing the variance of the degree distribution.

Using (8) and (9) we get

$$\begin{aligned}
S_N^2 &= \frac{2}{N}\left(T_{TS}+3T_T\right) \\
&\quad -(N-1)\left(\frac{4T_{OE}+8T_{TS}+12T_T}{N(N-1)(N-2)}\right)\left[1-(N-1)\left(\frac{4T_{OE}+8T_{TS}+12T_T}{N(N-1)(N-2)}\right)\right]
\end{aligned}$$

Inducing a mean-preserving spread requires triad manipulations that (i) increases the first term in the expression above, while (ii) leaving the second term unchanged. Table 3 list several mean-preserving triad manipulations. A triad is the smallest subgraph we can use to induce a mean-preserving spread in the degree distribution.

To increase S_N a two-star or triangle must be added to the network (accommodated by changes in the number of empties and one edges). Alternatively we can convert a two-star into a triangle (again accommodated by changes in the number of empties and one edges). These correspond to manipulations 2 to 5 and manipulation 6 in Table 3. Note that manipulations 2 and 4 and 3 and 5 are isomorphic, while manipulation 1 does not increase S_N^2. This leaves 4, 5 and 6 as unique triad manipulations which induce mean preserving spreads in a network's degree distribution.

Each of manipulations 4 to 6 involve net increases in $T_{TS}+3T_T$, accommodated by decreases in the number of one edges and increases in the number of empties. This is an example of

Table 3: Mean-preserving spreads via triad manipulations

		Net final change in triad type				Change in S_N^2	Change in R_N
#	Initial triad manipulation	empty	one edge	two star	triangle		
1	empty to one edge	0	0	0	0	0	0
2	empty to two star	+1	-2	+1	0	$\frac{2}{N}$	-
3	empty to triangle	+2	-3	0	+1	$\frac{6}{N}$	+
4	one edge to two star	+1	-2	+1	0	$\frac{2}{N}$	-
5	one edge to triangle	+2	-3	0	+1	$\frac{6}{N}$	+
6	two star to triangle	+1	-1	-1	+1	$\frac{4}{N}$	+

how a network's degree distribution fundamentally constrains other aspects of its topology. In this case higher variance degree sequences imply networks with more "hubs" (nodes with many links emanating outwards from them) *as well as* more isolated nodes. Jackson and Rogers (2007) show that the first effect tends to facilitate the spread of infections (ideas, new technology, etc.) in a network, while the second acts as a break to diffusion.

Note that the effect of the triad manipulations listed in Table 3 on transitivity is not one-directional. Manipulation 4 reduces transitivity, while manipulations 5 and 6 increase it.

References

[1] Albert, Reka and Albert-Lászlo Barabási. (2002). "Statistical mechanics of complex networks," *Review of Modern Physics* 74 (1): 47 - 97.

[2] Barabási, Albert-László and Réka Albert. (1999). "Emergence of scaling in random networks," *Science* 286 (5439): 509 - 512.

[3] Bonacich, Phillip. (1972). "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology* 2 (1): 113 - 120.

[4] Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. (2009). "Power-law distributions in empirical data," *SIAM Review* 51 (4): 661 – 703.

[5] Comola, Marghertia and Marcel Fafchamps. (forthcoming). "Testing unilateral and bilateral link formation," *Economic Journal.*

[6] Debreu, Gerard and I. N. Herstein. (1953). "Nonnegative square matrices," *Econometrica* 21 (4): 597 - 607.

[7] De Weerdt, Joachim. (2004). "Risk-sharing and endogenous network formation," *Insurance Against Poverty*: 197 - 216 (Dercon, Stefan, Ed.). Oxford: Oxford University Press.

[8] Jackson, Matthew O. (2008). *Social and Economic Networks.* Princeton, NJ: Princeton University Press.

[9] Jackson, Matthew O. (2014). "Networks and the identification of economic behaviors," *Mimeo, Stanford University.*

[10] Jackson, Matthew O. and Brian W. Rogers. (2007). "Relating network structure to diffusion properties through stochastic dominance," *B.E. Journal of Theoretical Economics* 7 (1) (Advances), Article 6.

[11] Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data.* New York: Springer.

[12] Milgram, Stanley (1967). "The small-world problem," *Psychology Today* 1 (1): 61 - 67.

[13] Newman, Mark E. J. (2003). "The structure and function of complex network," *SIAM Review* 45 (2): 167 – 256.